Integrating the signatures of demic expansion and archaic introgression in studies of human population genomics
Lauren Alpert Sugden\textsuperscript{1,2} and Sohini Ramachandran\textsuperscript{1,2}

Human population genomic studies have repeatedly observed a decrease in heterozygosity and an increase in linkage disequilibrium with geographic distance from Africa. While multiple demographic models can generate these patterns, many studies invoke the serial founder effect model, in which populations expand from a single origin and each new population’s founders represent a subset of genetic variation in the previous population. The model assumes no admixture with archaic hominins, however, recent studies have identified loci in Homo sapiens bearing signatures of archaic introgression. These results appear to contradict the validity of analyses invoking the serial founder effect model, but we show these two perspectives are compatible. We also propose using the serial founder effect model as a null model for determining the signature of archaic admixture in modern human genomes at different geographic and genomic scales.

Addresses
\textsuperscript{1}Center for Computational Molecular Biology, Brown University, Providence, RI, USA
\textsuperscript{2}Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA

Corresponding author: Ramachandran, Sohini
(sramachandran@brown.edu)

Introduction
Research in human evolution relies on multiple fields — such as archaeology, genetics, and linguistics — to give a history of Homo sapiens during the last hundred thousand years. In the last decade, investigations of worldwide human genomic variation based on multiple genetic marker types have observed three robust trends in summary statistics as a function of increasing geographic distance from Africa (Figure 1; see also DeGiorgio et al. [1]): a decrease in heterozygosity, an increase in linkage disequilibrium (LD), and an increase in the frequency of derived alleles [2–8]. Analyses of genomic datasets from human pathogens and parasites have also observed a reduction in heterozygosity as a function of distance from Africa [9–11]. Further, studies of anthropometric [12,13], economic [14], cultural [15], and linguistic data [16] have reported that population-level diversity in various traits follows this same pattern.

These studies each interpreted the observed reduction in trait diversity out of Africa as reflecting the population and geographic expansion of modern humans in the last 45k–60k years by invoking a model for the great human expansion known as the ‘serial founder effect model’ [2,3,7,17–20]. In a serial founder effect model, a series of successive bottlenecks from a single origin of expansion produces a stepwise increase in genetic drift — and consequent decrease in genetic diversity — as a function of geographic distance from the origin. Although the ability to identify a unique model for human evolutionary history from genome-wide diversity patterns has generated debate across a range of disciplines (for example, [1,21–29,30*,31,32*]), the serial founder effect model has proven to be a useful framework for both understanding the dynamic nonequilibrium history of human populations and identifying origins of large-scale human population expansions [6,19,33–35]. However, emerging high-quality archaic genomes and genomic datasets sampling increasing numbers of individuals from diverse human populations have shed new light on the complex demographic events that characterize human evolutionary history. Thus, the time is ripe to examine where the assumptions of the serial founder effect model falter, and what we can learn from genomic regions and populations in which genomic diversity patterns diverge from those predicted by the model.

We focus here on one particular assumption: the serial founder effect model, as invoked by Prugnolle et al. [2] and Ramachandran et al. [3], implicitly assumes that admixture with archaic hominins did not play a substantial role in the evolutionary history of H. sapiens (see also DeGiorgio et al. [1], Pickrell and Reich [29]). Here, we review observed patterns of human genomic variation from large-scale studies in the last five years, examine this assumption closely in light of these empirical results, and propose a framework for disentangling the signatures of archaic introgression from those of modern human population interactions in worldwide human genomic datasets.
Observed genome-wide patterns of heterozygosity and linkage disequilibrium (LD) among worldwide human populations as functions of geographic distance from Addis Ababa, Africa (9 N, 38 E). Haplotype heterozygosity [panel (a), calculated as in Conrad et al. [99]] and average LD at 10 kb [panel (b), measured by r²] were calculated across 640,034 SNPs genotyped in the Human Genome Diversity Panel [7]. Note error bars are smaller than symbols. Equations for fitted lines and each linear model’s coefficient of determination are displayed within each panel. Symbols indicate geographic regions: red circles = Africa, orange left-pointing triangles = Middle East, yellow squares = Europe, green down-pointing triangles = Central/South and East Asia, blue right-pointing triangles = Oceania, purple up-pointing triangles = Americas; populations are assigned to regions as in Rosenberg et al. [38].

**Figure 1**

**Human genomic variation reveals intercontinental clusters and intracontinental admixture**

Analyses of human genomic data from globally distributed populations repeatedly find that multilocus genotypes from human populations produce genetic clusters largely corresponding to major geographic regions [36–41]. Within continents, inferred genetic clusters may identify geographically or culturally isolated populations, distinguish among various subsistence strategies, or reveal signatures of gene flow [8,42,43].

The existence of genetic clusters among worldwide human populations has been challenged for multiple reasons [44–46]. One point of contention regarding the ‘clusteredness’ of humans is particularly relevant to this review: Serre and Paabo [47] questioned whether the identification of clusters from human multilocus genotype data is an artefact due to the sampling design of datasets such as the Human Genome Diversity Panel [4,5,37,48]. In response, Rosenberg et al. [38] showed that inferred clusters arise not from the geographic dispersion of sampled individuals, but rather are generated by small discontinuous jumps in genetic distance for population pairs on opposite sides of geographic barriers (Figure 2 and [44]).

The serial founder effect model assumes no gene flow between neighboring populations, and generates a positive correlation between genetic distance and geographic distance among population pairs [3]. As was observed by Rosenberg et al. [38] using microsatellite data, we find that the effect of a barrier that delineates a continental region — such as a large mountain range or a continental shelf — is to add to pairwise genetic distance between populations beyond the value predicted by geographic distance alone. In the case of HGDP SNP data analyzed in Figure 2, crossing a barrier adds 0.0703 to pairwise FST, the equivalent to traveling approximately 9764 km within a continental region.

Thus, Figure 2 strongly suggests that qualitative features of the serial founder effect model hold true for human evolutionary history: modern human genetic diversity generally reflects a stepwise accumulation of genetic drift across geographic barriers, and intercontinental migration was not
Genetic differentiation among worldwide human populations reflects intercontinental clusters and intracontinental clines. Pairwise genome-wide \( F_{ST} \) was calculated across various population pairs in the HGDP. Different colors represent comparisons within and between continental groups defined by major representation in the same cluster in the \( K = 5 \) plot of Figure 2 in Rosenberg et al. [38]. Green +’s indicate comparisons within each continental group (Africa, Eurasia, East Asia, Oceania, and America). ‘Eurasia’ here combines Europe, the Middle East, and Central/South Asia. Orange ‘x’s indicate pairs with one population from Eurasia and one from East Asia; purple ‘x’s indicate pairs with one population from Africa and the other from Eurasia; and pink circles indicate pairs with one population from East Asia and the other from either Oceania or America. Comparisons with the Hazara or Uyghur are indicated by an ‘H’ or ‘U’, respectively; these admixed populations are less differentiated from Eurasian populations than predicted by geographic distance alone. A linear regression of \( F_{ST} \) onto distance and geological barriers estimates \( F_{ST} = 0.0012 + 0.0072 \times D + 0.0703 \times B \), where \( D \) is great circle distance measured in thousands of km, and \( B \) is a Boolean variable indicating whether at least one barrier such as oceans or the Himalayas occur between a given population pair [38]. Both regression coefficients are significantly different from 0 (\( p < 0.001 \); \( R^2 = 0.773 \). Regression analyses reflecting the same qualitative pattern were generated by Rosenberg et al. [38] using microsatellite data in the HGDP.

The norm in human evolutionary history. This is borne out by analyses of diverse human populations in conjunction with the HGDP [6–8,34,35,49,50]. Further, the bottlenecks in the serial founder effect model, which increase drift and short-range linkage disequilibrium [51], is supported by studies that infer parameters for human demography, all of which identify bottlenecks during the peopling of new continental regions [52–55]. Indeed, models that attempt to recreate genomic patterns of LD and heterozygosity without bottlenecks must invoke either very carefully constructed migration rates [29] or an instantaneous divergence model with monotonically decreasing population sizes [1].

In contrast to observed intercontinental clusters, analyses of continental human genomic variation indicate limited population structure within continents (see Novembre et al. [56]), with counterexamples corresponding to unique population histories and waves of migration (for example, in sub-Saharan Africa [8], the Americas [43,49], and India [57]). However, two populations which notably diverge from the expected increase of genetic distance with geographic distance shown in Figure 2 are the Hazara and Uyghur, admixed populations deriving equal ancestry from Eurasian and East Asian genetic clusters inferred by Rosenberg et al. [38]. This deviation reflects the genomic signature of modern human admixture across geographic barriers, one of the situations in which the assumptions of the serial founder effect model are violated. Next we review empirical genomic patterns of archaic admixture with modern humans, which also deviate from patterns predicted by the model.

Genomic signatures of archaic admixture in \( H. \) sapiens

After years of debate regarding the contribution of archaic hominins to the modern human gene pool, studies taking advantage of recent advances in the amplification of DNA from ancient human remains have produced compelling evidence that archaic admixture occurred with ancestral modern humans. Further studies have shown that some of these introgression events were important for adaptation within modern human populations (see Racimo et al. [58] for a review). Although archaic hominin mitochondrial and Y-chromosomal haplotypes are not present among any modern humans [59,60], several studies of modern human populations outside of Africa infer genome-wide levels of archaic admixture ranging from 1% to 8% originating from more than one admixture pulse [61–63,64,65,66]. When combined with a serial founder effect in simulation, these inferred levels of archaic admixture produce genome-wide patterns in summary statistics like those observed in modern human data and unique patterns of population structure at a small fraction of loci in the human genome. In contrast, archaic admixture of 10% or greater both inflates measures of linkage disequilibrium and leads to a discontinuous relationship between heterozygosity and geographic distance [1].
Three genomic signatures are used to localize regions of archaic introgression in modern human genomes: identifying firstly, loci where non-African modern humans carry a derived single-nucleotide variant seen in archaic hominins but not in Africans [61,66]; secondly, non-African haplotypes with low sequence divergence to an archaic haplotype but high sequence divergence to African haplotypes [67]; and thirdly, candidate haplotypes bearing the two previous signatures with lengths consistent with interbreeding between 40 and 90 kya [68,69]. Loci meeting these criteria can then be tested for signatures of adaptive or balancing selection; to date, candidate genes for adaptive archaic introgression belong to skin pigmentation, immune, and metabolic pathways [58,61,67,70–73,74**,75,76*,77].

While adaptive introgression has been shown for select genomic regions, individual maps of archaic ancestry in modern humans suggest that introgression also was associated with deleterious fitness consequences for modern humans [78]. For example, Sankararaman et al. [75] find that genic regions exhibit reduced Neandertal ancestry, implying that modern human genetic lineages bearing archaic sequences have been removed by purifying selection over time.

Our understanding of the relationship between ancestral modern humans and archaic hominins is limited by the small number of high-quality genomes available from archaic individuals. Although technology for extracting whole genomes from ancient hominin and human remains is improving at a fast pace, there are regions of the world that may never be represented in genomic studies of archaic hominin or ancestral modern human remains, due to unfavorable climates for DNA preservation in fossils. The recent development of methods for constructing individual maps of archaic ancestry from modern human population-genomic data and reference archaic genomes is thus a promising approach to gain insight into hominin evolutionary history, and one with power to exhume population-genomic information from extinct hominins without a large collection of archaic sequences [64**,65*,75,79*].

**Leveraging the serial founder effect model in scans for archaic introgression**

While the genome-wide patterns in Figure 1 are well-predicted by the serial founder effect model, population genomic data are shaped by a plethora of evolutionary events beyond bottlenecks: including, but not limited to, population expansions, admixture (with archaic hominins and among modern humans alone), slavery, isolation, and adaptation [80]. Each of these events leaves its own genomic footprint, which might be either distributed genome-wide (e.g. an excess of rare variants genome-wide due to post-agricultural population expansions [81]) or more confined to a local genomic region (e.g. support from X-chromosomal variation for male-biased European ancestry in African Americans [42]). Here we suggest an integrative framework for reconciling the coexistence of genomic signatures of archaic admixture with genomic signatures of modern human demographic processes: by using the robust genome-wide patterns present in intercontinental human population genome datasets and predicted by the serial founder effect model as a null model, we can identify local genomic regions deviating from those patterns in predictable ways that indicate gene flow from archaic hominins (Figure 3).

Archaic introgression at a given locus leads to an increase in LD at that locus relative to genome-wide levels of LD predicted by the serial founder effect model (Figures 1b, 4a,b, and [58]). In Figure 4a, the East Asian populations in the HGDP show a clear inflation in LD at the OAS gene cluster involved in immune response to viral infection; Mendez et al. [72] identified ~185 kb introgressed from Neandertals into present-day non-Africans (particularly in Oceanic individuals and Indonesians) in this gene cluster. Arguably the most celebrated example of archaic introgression occurred from the Denisovan lineage into Tibetans at the gene *EPAS1*, which encodes a transcription factor affecting response to hypoxia at high altitudes (Figure 4b). Analyses of a single genome from the southern Chinese Dai population did not reveal either the Denisovan *EPAS1* haplotype or excess allele sharing with the Denisovan genome [63,74**]; Huerta-Sánchez et al. [74**] further note that three single-nucleotide variants in *EPAS1* are private alleles shared only between Tibetans and the archaic Denisovan genome, and that the putatively beneficial *EPAS1* haplotype is not observed in Melanesian modern humans or in the Altai Neandertal from Siberia [82*]. However, analysis of population genomic data has found higher Denisovan ancestry in southern Chinese populations — such as the Dai, highlighted in Figure 4b — than in northern Chinese populations [83].

Other evolutionary forces, notably selection in the modern human lineage, can also lead to a local genomic increase in LD relative to the genome-wide pattern. Strong positive selection, for example, should lead to an increase in LD for the population or populations with the adaptive allele. In Figure 4c, we observe inflation in LD worldwide in the highly conserved gene *FOXP2*, a transcription factor crucial for language development [84] and part of an 11 Mb introgression desert [65*]. In fact, Kuhlwim et al. [85**] classify a 150 kb segment within *FOXP2* as exhibiting signatures of gene flow from modern humans into the Altai Neandertal (see also Coop et al. [86]). Conservation through strong purifying selection is expected to inflate measures of LD [87], but will produce a different genomic signature than introgression: many human populations, including those in Africa, should be affected to a similar degree (Figure 4c), in contrast to the local genomic
signatures expected under positive selection confined to a particular geographic region (Figure 4a,b).

In Figure 4, we illustrate the utility of comparing patterns predicted by the serial founder effect model to local genomic patterns at known targets of archaic introgression and conservation by a simple comparison to a linear model generated by genome-wide data. Similar comparison of local genomic patterns to genome-wide patterns are often explored to determine the relative roles of hard adaptive sweeps versus demographic processes in shaping modern human genomic variation [88–90]. As more diverse human populations and archaic samples are sequenced at higher coverage, identifying loci undergoing archaic admixture will be done with greater precision. New methods will undoubtedly be developed that combine multiple statistics and multivariate outlier detection for this purpose (as was done to detect positive and balancing selection by Hunter-Zinek and Clark [91]). We suggest that the framework in Figures 3 and 4 outlines a fruitful avenue for future research, allowing investigators to disentangle the genome-wide footprints of modern human demographic history from evolutionary forces such as archaic admixture, selection, and conservation that act on local genomic targets.

**Conclusions and future directions**

DNA from human and archaic hominin remains offers unique snapshots into the evolutionary history of our species. However, our ability to integrate this data type into analyses of present-day human population genomic data is hampered by — and may always be hampered by — the limited number of archaic hominin samples with high-quality genomic data. More work needs to be done to evaluate how the systematic dearth of archaic samples in certain parts of the world biases downstream inferences. Additionally, the interaction between genetic drift in modern humans and pulses of archaic admixture has been underexplored (but see Skoglund and Jakobsson [83]). Therefore, approaches to identify the genomic signatures of archaic admixture from present-day human genomes, such as those developed by Vernot and Akey [64**], Vernot et al. [65*], Sankararaman et al. [75*],79*], will be crucial to inferring hominin evolutionary history. Here we propose one approach for identifying targets of archaic introgression in modern human genomes that leverages the robust genome-wide patterns predicted by the serial founder effect model, and scans for regions that diverge from these genome-wide patterns. Further evaluation and development of our framework — particularly in the form of methods for determining the
Figure 4

Linkage disequilibrium (LD) in loci undergoing archaic introgression and various forms of selection deviate from the patterns predicted by a serial founder effect model. LD is calculated among SNP pairs 10kb apart as in [4], and error bars are standard errors of mean LD across all SNP pairs within each analyzed genomic region. Symbols match those in Figure 1; the regression line fitted in Figure 1b is shown as a dashed black line in each panel with 99% confidence bands shaded in gray. (a) The OAS gene cluster (chr12:113300000–113600000 (hg19)), shown to have Neanderthal gene-flow in non-Africans [72], shows a deviation from genome-wide patterns, particularly in East Asia (green down-pointing triangles). (b) EPAS1 (chr2:46524563–46613836), a gene crucial for high-altitude adaptation in Tibetans [96] and known to bear an introgressed Neanderthal haplotype [74*], deviates from genome-wide patterns in the Dai population, a southern Chinese ethnic minority group. There are no Tibetan samples in the HGDP. (c) 11 Mb region surrounding FOXP2 (chr7:113600000–124700000), previously identified as an introgression desert [65*]. This genomic region has elevated LD worldwide relative to genome-wide patterns (grey dashed regression line), and all populations seem to be equally affected, which may be a signature of background selection/conservation.

genome-wide significance of local genomic patterns of population-level heterozygosities — will be necessary before this framework can yield statistically sound inferences; still, we believe that the proposed framework will ultimately be a versatile and fruitful one. For example, future studies could use simulations to learn expected patterns of genome-wide versus local genomic diversity under a range of models that violate the assumptions of the serial founder effect: adaptive selection, super-exponential growth, and various pulses of archaic admixture (as studied by [65*,72,79*,92*], who aim to differentiate between signatures of Neanderthal versus Denisovan admixture).

The framework we propose also underscores the fact that the relative role of archaic admixture in shaping present-day human genetic variation is small when compared to the genomic footprint of modern human population
Acknowledgements
We thank Trevor Pemberton, Sitram Sankararaman, and Carina Schlebusch for helpful discussions, Arthur Sugden for help generating Figure 3, and John Novembre for comments on an earlier draft of this manuscript. This work was supported by the Pew Charitable Trusts (SR is a Pew Scholar in the Biomedical Sciences), and US National Institutes of Health (NIH) grant R01GM118652 (to SR) and COBRE award P20GM109035. SR also acknowledges support from NSF CAREER Award DBI-1452622.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest


22. Dahl O: Are small languages more or less complex than big ones? Linguist Typlol 2011, 15:171-175.


This study provides new conceptual and methodological approaches for studies of genetic and linguistic coevolution, rejects the serial founder effect model for sound units, and clarifies the different population processes that affect genes and languages at the global versus regional geographic levels.


Through analysis of an expanded genetic dataset, the authors reject the claim made by Ashraf and Galor [14] that genetic diversity has driven comparative economic development.


This study analyses a Neandertal Y chromosome from Spain and places the lineage as a distant outgroup to modern human Y chromosomes; the estimated time to the most recent common ancestor between this Neandertal lineage and modern human Y chromosomes is over 580k years ago.


64. Vernet B, Akey JM: Resurrecting surviving Neandertal lineages. Science 2014, 343:1017-1021. The authors apply a two-stage computational approach to modern human genomes to identify population-level DNA sequence information from Neandertals, and infer a per-individual average of approximately 19.6 Mb of Neandertal sequences in over 600 European and East Asian individuals. This study also confirms that Neandertal admixture occurred multiple times with ancestral modern humans outside of Africa (see also [92,77,65]).


The authors show that the adaptation of Tibetans to the hypoxic environment of the Tibetan plateau can only be explained by the introgression of a Denisovan haplotype around the gene EPAS1. This study suggests archaic admixture may have helped ancestral modern humans adapt to new environments during the modern human expansion out of Africa.


This study identifies introgression deserts and illuminates the role purifying selection has played in removing genetic material from Neandertals in modern human genomes; findings suggest that the X chromosome played an important role in reproductively isolating modern humans and archaic hominins. The authors also report multiple Neandertal-derived alleles that increase disease risk in modern humans.


This study by the SIGMA Type 2 Diabetes Consortium identifies a new risk haplotype for type 2 diabetes spanning two sole carriers (SLC16A1 and SLC16A13) that introgressed into modern humans from Neandertals.


The authors unravel segments of Denisovan and Neandertal ancestry in 120 diverse modern human populations, and report more Denisovan ancestry in South Asians than is predicted by current models of archaic admixture.


82. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C et al.: The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 2014, 505:43-49. This study reports a high-quality genome sequenced from a Neandertal woman found in Siberia, and infers several gene flow events in the late Pleistocene among archaic hominins and between archaic hominins and modern humans. One of these gene flow events indicates the Denisovan lineage may derive ancestry from a currently unknown archaic group.


This study jointly analyses whole genomes from archaic and modern humans and is the first to characterize gene flow from modern humans into Neandertals, in this case into Neandertals from the Altai Mountains of Siberia. The authors do not find evidence for gene flow from modern humans into the Denisovan or European Neandertal genomes.


The authors show that genetic drift cannot explain the observed higher proportion of Neandertal ancestry in East Asians compared to Europeans. Using simulations, this study find strong support for multiple pulses of Neandertal admixture in modern human genetic lineages (as observed by [64]).


